On the relative importance of audio and video in the presence of packet losses

Korhonen, Jari; Reiter, Ulrich; Myakotnykh, Eugene

Published in:
proceedings QoMEX

Publication date:
2010

Document Version
Publisher's PDF, also known as Version of record

ON THE RELATIVE IMPORTANCE OF AUDIO AND VIDEO IN THE PRESENCE OF PACKET LOSSES

Jari Korhonen[1]

Department of Photonics Engineering
Technical Univ. of Denmark (DTU)
Lyngby, Denmark

Ulrich Reiter, Eugene Myakotnykh

Centre for Quantifiable Quality of Service
in Communication Systems (Q2S)[2]
Norwegian Univ. of Science and Tech. (NTNU)
Trondheim, Norway

ABSTRACT

In streaming applications, unequal protection of audio and video tracks may be necessary to maintain the optimal perceived overall quality. For this purpose, the application should be aware of the relative importance of audio and video in an audiovisual sequence. In this paper, we propose a subjective test arrangement for finding the optimal trade-off between subjective audio and video qualities in situations when it is not possible to have perfect quality for both modalities concurrently. Our results show that content poses a significant impact on the preferred compromise between audio and video quality, but also that the currently used classification criteria for content are not sufficient to predict the users' preference.

Index Terms — Subjective quality assessment, Video quality, Audio quality, Packet loss stream

1. INTRODUCTION

In multimedia streaming over packet-switched networks, packets are typically protected against losses by using either *forward error correction* (FEC), retransmissions, or both. Bandwidth restrictions and packet deadlines set an upper limit for the proportion of the link capacity that can be used for recovering the lost packets. If these restrictions do not allow recovering all the lost packets, it may be useful to apply *unequal error protection* (UEP), where the most important packets have the strongest protection. UEP can be implemented by allocating more redundancy or a higher number of retransmission attempts to the high priority packets.

Unfortunately, it is not straightforward to define the relative priority of data packets. Several different approaches have been proposed for packet classification in video and audio streaming [1,2], but less effort has been put into analyzing the relative importance of video and audio streams when they both are present. That there is also a cross-modal influence of perceived quality between audio and video has been shown in [11]. Some joint quality models have been proposed to measure the overall quality when the subjective audio and video qualities are known separately (see [3], for example). Even though rough overall quality estimates can be obtained from these models, many studies show substantial variation in the results, depending on the context and content type [3,4]. This is why further investigations are needed to improve joint audiovisual quality assessments in different scenarios, and pave the way for a more accurate objective metric.

In this paper, we propose a novel method for identifying the optimal subjective overall quality within given constraints for audio and video distortions. More specifically, we focus on the situation, where a predefined proportion of lost packets must be balanced between audio and video streams. In our method, test subjects are presented with synchronized audio and video sequences. Subjects are asked to adjust the balance between audio and video *packet loss rate* (PLR) by moving a slider. In one extreme, all packet losses occur in the video stream, with the audio stream being free of errors. In the opposite extreme it is the other way round. The task of the test subjects is to indicate the perceptually optimal trade-off between audio and video qualities.

Even though there are only few practical applications today in which a straightforward trade-off between audio and video quality is required, such a scenario can be relevant in multimedia communications, as we will demonstrate in Section 2. Furthermore, our study improves the general knowledge about the relative importance of audio and video tracks in communications applications, and we believe that such knowledge is useful for both, application development, and content creation.

The remainder of this paper is organized as follows. In Section 2, we briefly review the relevant related work regarding subjective quality assessment and UEP. In

[1] This work was done during employment at Q2S, NTNU.

[2] "Centre for Quantifiable Quality of Service in Communication Systems, Centre of Excellence" appointed by the Research Council of Norway, funded by the Research Council, NTNU and UNINETT.

Section 3, we formulate the problem statement more specifically, and the proposed method and test software are explained in detail. In Section 4, the results are presented and discussed. Finally, Section 5 provides conclusions and an outlook.

2. BACKGROUND

2.1 Quality assessment

Many different objective (i.e. algorithm based) quality metrics have been proposed for audio and video quality assessment. However, in spite of the progress in objective quality assessment algorithms, subjective quality assessments involving human subjects are still considered the ultimate way of quality evaluation. There are several standardized methods to perform subjective quality assessments. Most of these involve some kind of rating scale, where test subjects are asked to rate the quality of the test sequence, often using a given anchor sequence as a reference point (double stimulus). Rating can be based on a binary scale (acceptable / unacceptable) or *mean opinion score* (MOS).

In the presence of both audio and video, the co-impact of audio and video quality distortions must be considered to assess the overall perceptual experience. The most straightforward method is to assess video and audio quality separately, and then combine these metrics into an overall quality metric. Several studies show that for certain applications and usage scenarios, the overall perceived quality can be reasonably well modeled by combining subjective video and audio MOS values using simple multiplicative, additive, or bilinear models. However, the weighting factors for audio and video are different, depending e.g. on content: in [3,4] it is suggested that in case of intensive motion, video quality should be given more weight than audio quality. It is also pointed out in [3] that most of the studies in this field focus on source distortion (=compression artifacts), or have been performed under rather artificial test conditions. Apparently, more research is needed to study models for realistic scenarios, involving both source and channel distortions.

A major problem with MOS scales in subjective quality assessments is that the results are strongly influenced by different factors, such as the vocabulary used for the scale, and training of test subjects [5]. This is why MOS results obtained from different studies are not necessarily comparable with each other. In practical scenarios, the intention is usually to compare two (or more) test objects and find their relative quality levels (i.e. which test sequence has the best quality), rather than MOS values describing the (often quite meaningless) absolute quality level. This aspect is emphasized when sequences with different content or different distortion types are compared.

In our earlier work [6], we have developed a subjective test method for comparing video sequences with different types of distortions. Unlike traditional subjective quality assessment methods with double stimulus, in our method test subjects are shown an anchor sequence with fixed distortion level, and a test sequence with an adjustable level of different type of distortion. The task of the test subject is to adjust the distortion level of the test sequence such that the perceptual qualities of the two sequences match each other as closely as possible. This method can be used to determine perceptually equivalent distortion levels for different types of artifacts.

In this paper, we extend the idea of turning incomparable into comparable, by taking two modalities into account, namely video and audio. In our test arrangement, video and audio qualities are inversely dependent on each other, and the task of the test subject is to find the perceptually optimal compromise. We assume that the results of practical studies employing the method will be useful for joint optimization of audio and video qualities in streaming applications suffering from transmission errors.

2.2 Unequal protection of streams

In this paper, we focus on an application scenario in which related audio and video streams are transported over a lossy packet network in parallel, and capacity constraints do not allow full recovery of all the lost packets. In this case, it may be necessary to protect audio and video streams unequally to obtain the best possible overall quality. There are several different alternatives to allocate uneven protection levels to transport streams with different relative priorities. One method is to employ selective retransmissions, so that the lost packets of highest priority are retransmitted first, and those of lower priority are retransmitted only if the remaining link capacity and delay constraints allow. Another possibility is to use unequal FEC code rates for streams with different priorities.

A common misunderstanding is that since the bit rate for a compressed audio stream is usually substantially lower than for the respective video stream, unequal protection of the two streams is a trivial issue. It is true that a low bit rate audio stream requires a smaller absolute amount of redundant data to obtain the same level of protection as a high bit rate video stream, but the question of optimal relative protection levels remains still open. If the video stream is perceptually much more important, it may be even beneficial to allocate the whole redundancy budget to video and leave audio unprotected.

Figure 1 demonstrates the impact of different relative protection levels on the residual packet loss rates in one possible example scenario. In this example, we assume that the bit rate for video is 2.5 times the bit rate for audio: there are 125 video packets and 50 audio packets per time unit to be protected separately by Reed-Solomon block codes. The

total redundancy budget is 10 packets per time unit, to be shared between audio and video streams. To simplify things, here we assume that the packets are equal in size. PLR in the channel is constant at 5%. The equations in [7] have been used to compute the theoretical residual PLR. Due to the space constraints, details are not explained here.

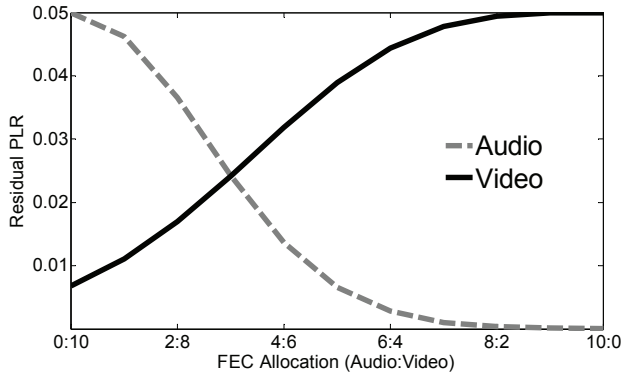

Figure 1. Residual PLRs for audio and video streams in a practical example scenario.

As can be seen from Figure 1, different trade-offs between residual PLRs for audio and video streams can be achieved by sharing the redundancy packets differently between audio and video streams. This is in spite of different bitrates of the streams. It should be noted that this observation is not universal; with different combinations of block lengths, redundancy overhead and channel PLR, curves for residual PLRs would be different. However, this example shows that scenarios *do exist* for which the loss rates for two streams are essentially inversely dependent on each other. In the remainder of this paper, and for the sake of simplicity, we assume that the sum of the residual PLRs for audio packets (PLR_{AUDIO}) and video packets (PLR_{VIDEO}) equals to the PLR in the channel ($PLR_{CHANNEL}$), as given in Eq. (1).

$$\mathrm{PLR}_{\mathrm{AUDIO}} + \mathrm{PLR}_{\mathrm{VIDEO}} = \mathrm{PLR}_{\mathrm{CHANNEL}} \qquad (1)$$

3. PROPOSED METHOD

3.1 Test methodology

The conceptual structure of the test software is illustrated in Figure 2. Due to performance limitations, the software used pre-processed raw video and audio files for playback. These had been generated by producing packet losses to encoded sequences and then decoding the lossy sequences. With respect to error concealment, frame repetition was used for audio sequences, and standard motion copy method as implemented in the H.264/AVC reference decoder (JM12.4) was used for video sequences. Test subjects moved a single slider to select the preferred pair of audio and video sequences, representing different trade-offs for audio and video distortion levels.

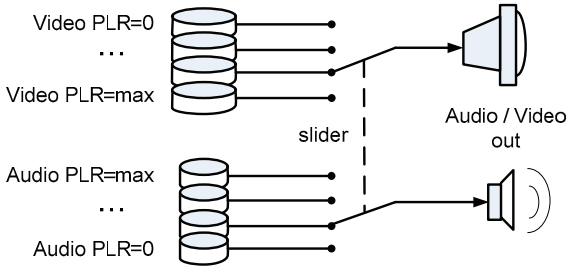

Figure 2. Conceptual diagram of the test system.

The interface designed for the subjective quality assessment study is shown in Figure 3. It included a window for playing the video sequences and a scale with no labels divided into ten equal intervals (subjects could only set the slider to discrete values on the scale). Audio was synchronized with video and played through an external sound card (EDIROL UA-25). Subjects were asked to adjust the balance between residual audio and video PLRs by moving the slider to the preferred position. On the left hand side of the scale, equivalent to scale point 1, video was at highest quality and all packet loss errors occurred in the audio domain. On the right hand side, at scale point 11, all packet losses occurred in the video stream, and the audio stream was free of errors. The sum of the residual loss rates was always equal to $PLR_{CHANNEL}$ (Eq. 1), which for our experiment was chosen to be equal to 2%, 5%, and 10%. Test subjects were asked to indicate the perceptually optimal trade-off between audio and video qualities for different PLRs and different content.

3.2 Test clips selection

Audiovisual samples representing different types of content (sports, music clip, cartoon, documentary/movie, news) have been used in the study. The selected audiovisual clips are characterized by different properties of audio and video. For the subjective experiment, seven samples were selected. The audio and video characteristics are described in Table 1.

Each clip contained a meaningful segment of audio and video information. The start and end points of the clips were selected such that semantic structures were maintained (e.g., complete sentences in the case of speech or singing). The selected test material represented different types of audio and video with different properties. The audio included speech, speech with background noise (crowd in a football stadium), singing (opera and pop music singing), and background music. The selected video clips had different spatial and temporal characteristics. The samples were taken from content as could be seen on TV.

Content	Video	Audio
Movie1	Details: high Motion: moderate	Speech, slow music
Movie2	Details: high Motion: high	Speech, fast music, sound effects
Opera	Details: moderate Motion: low	Opera singing (male)
Pop	Details: high Motion: high	Singing, pop music (female)
News	Details: moderate Motion: low	Speech
Football	Details: high Motion: high	Speech, crowd noise
Cartoon	Details: moderate Motion: moderate	Speech

Table 1: Selected samples.

Figure 3.User interface of the test software with slider position indicating preference of highest video quality.

3.3. Processing of test material

The original high quality, high resolution video sequences were decompressed, resized, and encoded in CIF resolution (352x288 pixels). CIF resolution was considered the most relevant for our study, since the target application is video streaming in a mobile environment. Here, CIF is the most widely supported format. To avoid quality fluctuations within clips, we have used constant *quantization parameter* (QP) instead of constant bit rate. This is why the bitrates for different compressed streams vary rather significantly, from 240 kbit/s ('Opera'), up to 1.3 Mbit/s ('Pop'). QP=30 was chosen to keep the source distortion level relatively close to unnoticeable. The *flexible macroblock ordering* (FMO) tool in checkerboard mode was enabled as an error resilience tool.

The audio streams were encoded by AAC codec with 128 kbps bit rate (2 channels). This codec and bit rate was chosen because a resulting encoded stream has a quality which is perceived as excellent by most people [10]. Two of the seven clips used in the experiment were monophonic ('News', 'Football', with the two channels being identical), whereas the other 5 contained stereo sound. The loudness level of audio in all clips was equalized according to their B-weighted signals. The experimental conditions are summarized in Table 2.

Number of samples	7
Video frame rate	24-30 fps
Video resolution	CIF (352 * 288 pixels)
Video color scheme	16 bit YUV (4:2:0)
Duration	30-40 seconds
Audio codec	AAC, 128 kbps, stereo, fullband
Video format	H.264/AVC, QP=30, GOP=10 frames, FMO
Max packet loss rates	2%, 5%, and 10 %, random distribution

Table 2: Test conditions used in the subjective study.

3.4 Test subjects and environment

A total of 40 subjects (32 male, 8 female) between 24 and 47 years of age (M=31.1, SD=4.94) participated in the experiment. Subjects were screened for color vision deficiency using Ishihara test charts, resulting in one subject with red-green blindness. All subjects reported normal or corrected to normal visual acuity. 37 subjects reported to have normal hearing, 3 subjects reported a doubt with respect to that before the experiment, and 2 of them actually slightly increased the playout levels for the experiment during the training. Therefore, during the assessment their playout level was around 2dB higher than for the remaining 38 subjects. This is considered uncritical for this experiment, as we are looking at packet loss artifacts (and not compression artifacts).

Of the 40 participants, 24 subjects had participated in earlier video quality assessments, and 15 in earlier audio quality assessments. A total of 11 subjects had attended both, audio and video quality assessments, before. They can be regarded as experienced assessors.

Subjects received detailed written instructions before the start of the experiment. A training session, which familiarized subjects with the test methodology, the type of distortions presented, and the test software's user interface, preceded the main part of the experiment. A fixed subset of the test material was used for training. Training included the

possibility for subjects to ask questions about the procedure before and after the training session. Training sessions lasted between 5 and 19 minutes (M=8, SD=3).

Subjects were allowed to adjust the playout level from the sound card during the training phase, but not during the actual experiment. They were provided with circumaural, open headphones (Beyerdynamic DT 990 PRO) for sound reproduction. Video content was displayed on a standard consumer 19" LCD computer monitor. The videos each had a size of approximately 11cm across on the monitor. Subjects were sitting at a viewing distance of 40cm. The background color (desktop color) was set to mid grey.

The experiment was performed in an acoustically treated room especially designed for audio and video quality tests. The wall behind the screen was uniformly illuminated by a daylight color temperature wallwash at 200 lux, with the remainder of the windowless room remaining with low illumination according to ITU-R BT.500-11 [8].

In the main section of the experiment, a total of 23 test items were presented. The first two items were considered to be warm-ups, for which no rating was recorded. The subsequent 21 test items were presented in random order for each subject, with no immediate repetition of content. The duration of the main section was between 12 and 40 minutes (M=25, SD=7). Hence, the average duration of the subjective assessment (including training) was around 33 minutes.

4. ANALYSIS AND RESULTS

The collected data was analyzed using descriptive and statistical analysis software. Except for 'Football' (all PLRs) and 'Movie2' (5% and 10% PLRs), distribution within each item (content at certain PLR) can be considered normal based on skewness and kurtosis analysis. Homogeneity of variance was assumed for all clips but 'Football', based on Levene's test.

The magnitude of impact of certain PLRs on perceived audio and video quality is largely unexplored. More specifically, it can be assumed that an equal distribution of packet loss errors between the audio and video parts of the transmission stream does not necessarily result in equally strong deteriorations of perceived quality in the two domains. Hence, the mean preference across all contents and PLRs was calculated to serve as a ground truth. On the 11-point preference scale, where 1 corresponds to error-free video and 11 to error-free audio, subjects preferred a mean value of 5.43 on average. It is important to note that this does not indicate a general preference towards higher video quality (rather than audio quality) in itself. The mean preference may vary for different contents and bit rates of the audiovisual stream.

The mean preference across the three PLRs tested (2%, 5%, and 10%) was found to be 5.27, 5.46, and 5.58, respectively. The slight tendency towards audio quality at increased PLRs was not significant: a post hoc Tukey$_b$ test showed that, with alpha at 0.05, means for all PLRs formed homogeneous subsets for all seven contents tested here. At the same time, standard deviation and variance increased slightly with higher PLRs, indicating a tendency that for subjects it was generally (but non-significantly) more difficult at higher PLRs to decide on a preferred modality for errors to appear in.

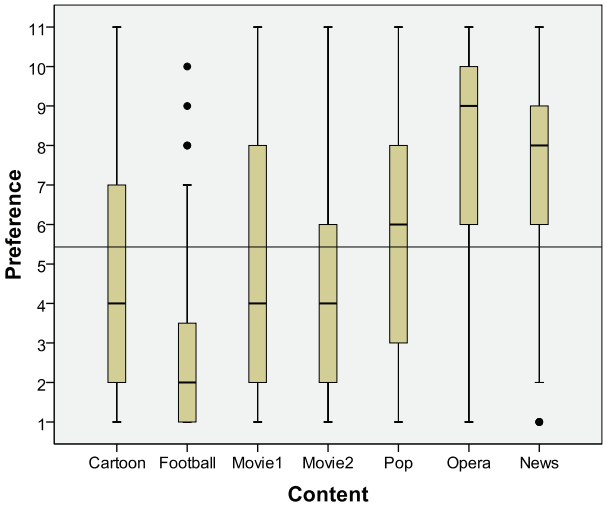

Figure 4. Box-plot of preferred trade-off, 1=best video quality, 11=best audio quality. Dots are outliers. Horizontal line=mean preference at 5.43.

Figure 4 shows a box-plot of the preferred trade-off between audio and video quality. As expected, distribution of preference across different content was not normal. This may indicate that the content used in this experiment was not necessarily representative. On the other hand, it is obvious that a classification of content can only be based on accepted classification criteria (like *detail* and *motion* in the video domain, see Table 1). From this we conclude that appropriate content classification criteria should be the topic of further research.

Looking at clips with similar *detail* and *motion* levels, one would expect that these received similar preference ratings, whereas clips with differences in these criteria would receive *different* preference ratings. Again, a post hoc Tukey$_b$ test showed that, with alpha at 0.05, means for content with similar criteria levels formed homogeneous subsets, see Table 3. This means that preference ratings for clips belonging to the same subset are not significantly different from one another. Surprisingly, 'Football' formed its own subset, although it was considered to have the same *detail* and *motion* levels as 'Movie2' and 'Pop'. This suggests that differences in the audio domain were larger

than can be expressed by only looking at the type of audio signal (*speech* / *non speech*, *music* / *non music*). At the same time, this also indicates the necessity for further research into cross-modal content classification criteria. This could be based on the work by Woszczyk et al. [9], who suggest the use of a matrix consisting of subjective attributes in different dimensions of perceptual experience, to be used in assessments of home theater systems. We also suspect that the emotional impact of different content plays a major role.

Homogeneous Subsets						
	Content		Subset for alpha = 0.05			
		N	1	2	3	4
Tukey$_b$	Football	120	2.79			
	Movie2	120		4.35		
	Cartoon	120		4.92	4.92	
	Movie1	120		5.04	5.04	
	Pop	120			5.67	
	News	120				7.30
	Opera	120				7.98

Table 3: Homogeneous subsets of preference.

The rather lengthy statistical analysis process involved in arriving at the homogeneous subsets as shown in Table 3 can be simplified if one is only interested in an approximated preference tendency Δp_i. This tendency can be expressed as the difference between the mean preference across all contents j and PLRs k and the mean preference for a specific content p_i (Eq. 2):

$$\Delta p_i = \overline{p_i} - \frac{1}{mno} \sum_{i,j,k=1}^{m,n,o} p_{ijk} \tag{2}$$

The strength S of this tendency can then be expressed as the absolute difference between the two, as given in Eq. 3:

$$S = \left| \Delta p_i \right| \tag{3}$$

Whether this tendency is significant or not, i.e. the tendency towards a preference of either good audio or good video quality is strong or not, can be tested by comparing the square root of the standard deviation with the strength S. The tendency can be considered strong whenever S is larger than the square root of the standard deviation of the preference p_i.

5. CONCLUSIONS AND FUTURE WORK

In this paper we have described a novel subjective test methodology for finding the optimal trade-off between subjective audio and video qualities. Such a method is useful in situations when it is not possible to have perfect quality for both modalities concurrently. Our results show that the content itself poses a significant impact on the preferred compromise between audio and video quality, and hence determines the relative importance of audio and video.

A successful estimation of an optimum audiovisual quality trade-off would therefore also require a differentiation of content in terms of relative importance of audio and video. We have shown that the commonly used classification criteria may not be sufficient for a useful differentiation. This is presumably the case because they frequently ignore the emotional aspects of semantically coherent content.

In our future work, we will therefore look into suitable classification criteria for audiovisual content. Here, we have considered the effect of channel distortion only. We also plan to extend our work to include source distortion, higher resolution video, and multichannel audio.

REFERENCES

[1] R. Hamzaoui, V. Stanković, Z. Xiong, "Optimized Error Protection of Scalable Bit Streams", *IEEE Signal Processing Magazine*, 22 (6), pp. 91-107, Nov. 2005.

[2] J.C. De Martin, "Source-driven Packet Marking for Speech Transmission over Differentiated-services Networks", Proc. of ICASSP'01, Salt Lake City, UT, USA, May 2001.

[3] D. Hands, "A Basic Multimedia Quality Model", *IEEE Trans. on Multimedia*, vol. 6., no 6., pp. 806-816, December 2004.

[4] S. Jumisko-Pyykkö, J. Häkkinen, G. Nyman, "Experienced Quality Factors: Qualitative Evaluation Approach to Audiovisual Quality", Proc. SPIE, Vol. 6507, San Jose, CA, USA, 2007.

[5] S. Zielinski, "On the Use of Graphic Scales in Modern Listening Tests", 123rd AES Convention, NY, USA, October 2007.

[6] U. Reiter, J. Korhonen, "Comparing Apples and Oranges: Subjective Quality Assessment of Streamed Video with Different Types of Distortion", Proc. of QoMEX'09, San Diego, CA, USA, July 2009.

[7] J. Korhonen, P. Frossard, "Flexible Forward Error Correction Codes with Application to Partial Media Data Recovery", *Signal Proc.: Image Comm.*, vol. 24, no. 3, pp. 229-242, March 2009.

[8] Recommendation ITU-R BT.500-11, "Methodology for the subjective assessment of the quality of television pictures", International Telecommunication Union, Geneva, Switzerland, 2002.

[9] W. Woszczyk, S. Bech, V. Hansen, "Interactions Between Audio-Visual Factors in a Home Theater System: Definition of Subjective Attributes", 99th AES Convention, NY, USA, 1995, Preprint 4133.

[10] D. Mears, K. Watanabe, E. Scheirer, "Report on the MPEG-2 AAC Stereo Verification Tests", ISO/IEC JTC1/SC29/WG11, N2006, 1998.

[11] J.G. Beerends, F.E. de Caluwe, "The influence of video quality on perceived audio quality and vice versa", J. Audio Eng. Soc., vol. 47, pp. 355-362, May 1999.